COMPUTATIONAL ANDSTRUCTURAL BIOTECHNOLOGY JOURNAL

Volume No: 5, Issue: 6, February 2013, e201302003, http://dx.doi.org/10.5936/csbj.201302003

CSBJ

Conformational Sampling in Template-Free Protein Loop Structure Modeling: An Overview

Yaohang Li [a,*]

Abstract: Accurately modeling protein loops is an important step to predict three-dimensional structures as well as to understand functions of many proteins. Because of their high flexibility, modeling the three-dimensional structures of loops is difficult and is usually treated as a "mini protein folding problem" under geometric constraints. In the past decade, there has been remarkable progress in template-free loop structure modeling due to advances of computational methods as well as stably increasing number of known structures available in PDB. This mini review provides an overview on the recent computational approaches for loop structure modeling. In particular, we focus on the approaches of sampling loop conformation space, which is a critical step to obtain high resolution models in template-free methods. We review the potential energy functions for loop modeling, loop buildup mechanisms to satisfy geometric constraints, and loop conformation sampling algorithms. The recent loop modeling results are also summarized.

Mini Review Article

Introduction

A loop, also called a coil, is a flexible segment of contiguous polypeptide chain that connects two secondary structure elements in a protein. The loop regions play critical roles in protein functions, such as involving in catalytic sites of enzymes [1], contributing to molecular recognition [2-4], and participating in ligand binding sites [5-7]. As a result, accurate prediction of the loop regions conformations in proteins is important for a variety of structural biology applications, including determining the surface loop regions in comparative modeling [8], defining segments in NMR spectroscopy experiments [9], designing antibodies [10], identifying function-associated motifs [11], and modeling the dynamics of ion channels [12, 13].

According to the loop length distribution illustrated in Figure 1, 93.2% of loops have lengths ranging from 2 to 16 residues, although sometimes loops can stretch much longer. Nevertheless, due to their high flexibility, loops regions are usually more difficult to model and analyze than the other secondary structures such as helices or strands. Indeed, in many (complete) protein models derived from computational methods, the loop regions, particularly the long ones, are the places contributing a lot of error [77]. At the early attempt of loop modeling, Flory [14] assumed that the backbone torsion angles corresponding to one residue are random, more precisely, statistically independent from the backbone torsions of its neighbors. However, more and more experimental [15], evolutional [16], and statistical [17] data have shown that loops are far from random and the nearby residue neighbors in sequence are sufficiently strong to account for substantial changes in the overall structure of loops. Figure 2 shows the ϕ-ψ propensity maps of Leucine in loops when the hydrophobic residues (ILE and VAL) are presented as neighbors at different distances. One can find that the backbone dihedral angle conformations of Leucine have strong correlation with the types of residues at the nearest and second nearest neighboring positions. However, such influences from residues at further positions are much weaker. The ϕ-ψ propensity maps of Leucine with ILE and VAL as two positions away neighbors are almost indistinguishable to the one of singlet Leucine, indicating that influences from neighboring loop residues two positions or further away are negligible. Moreover, studies have demonstrated that the identical peptide segments can adopt completely different structures in different proteins [18, 19]. Hence, in addition to the residues in a loop, the residues surrounding the loop structure are also important to determine its conformation, particularly for a loop deeply embedded in the protein structure. Furthermore, the distance between the anchor points in the rest of the protein that spans the loop likely influences the loop conformation as well, particularly when the loop is short. To facilitate studies on 3D structures of loops, the Protein Coil Library [20] maintains the structures of all loop segments derived from protein structures presented in Protein Data Banks (PDB).

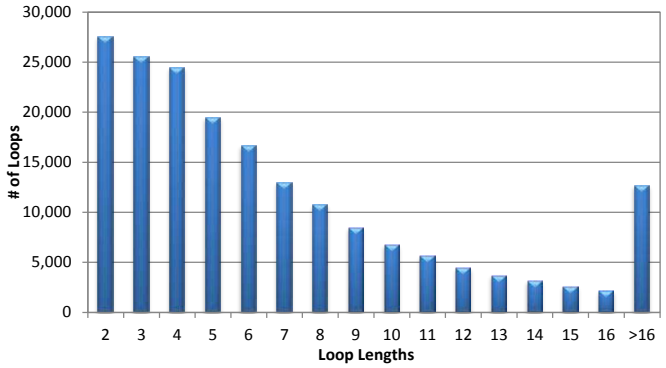

Figure 1. Distribution of loop lengths in the protein chain list generated by the PISCES server [21] on Aug. 28, 2012 containing 13255 chains with 2.0A resolution, 90% sequence identity, and 0.25 R-factor cutoff.

[a]*Department of Computer Science, Old Dominion University, Norfolk, VA 23529, USA*

* Corresponding author.
E-mail address: yaohang@cs.odu.edu

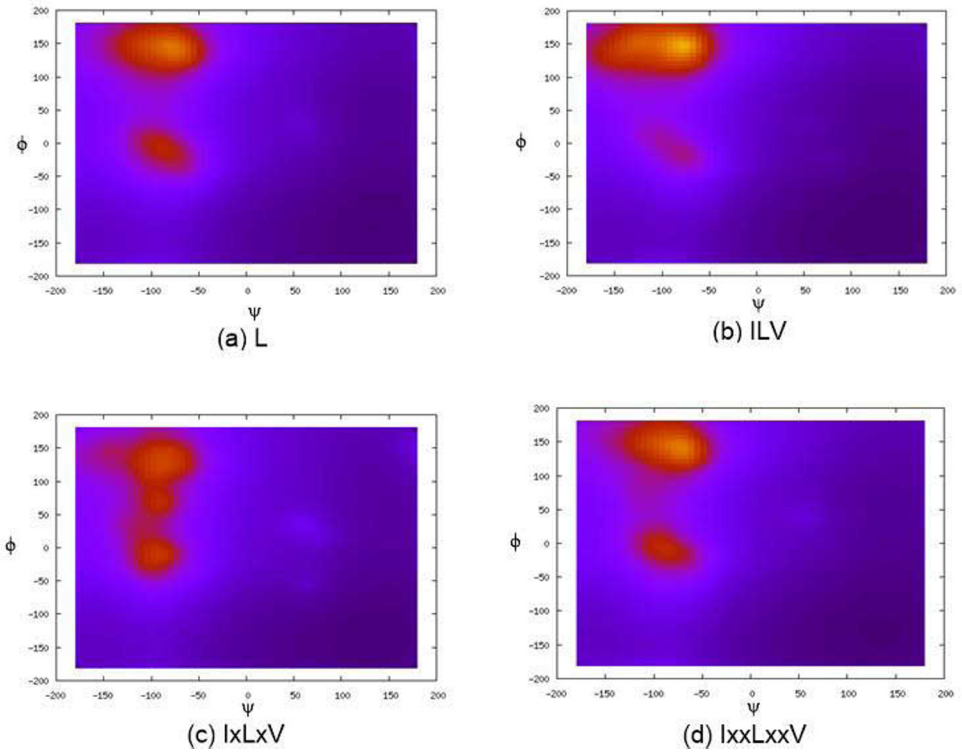

Figure 2. ϕ-ψ propensity maps of Leucine in the loops in presence of hydrophobic neighbors (ILE and VAL): (a) LEU as a singlet; (b, c, d) LEU with ILE and VAL as the nearest, one position away, and two positions away neighbors in sequence. The nearest and second nearest neighbors have strong influences to the backbone torsion angle conformations of Leucine and the influences from further neighbors are significantly weakened.

In general, loop structure modeling methods can be categorized into template-based (database search) methods and template-free (*ab initio*) methods. The template-based methods [22-25] search PDB for loop structure templates that fit the geometric and topologic constraints of the loop stems. The template-based methods highly depend on the quality and number of known structures in the PDB. Due to the fact that the number of possible loop conformations grows exponentially with lengths, the template-based methods are limited to relatively short loops. In contrast, the template-free methods can avoid this problem by sampling loop conformation space guided by energy functions. In this mini-review, we focus on the template-free methods only.

The template-free loop modeling problem is regarded as a "mini protein folding problem" [47] under geometric constraints, such as loop closure and avoidance of steric clashes with the remainder of the protein structure. Similar to the protein folding problem, modeling steps including coarse-grained sampling, filtering, clustering, fine-grained refining, and ranking are often found in most loop structure prediction methods. During the coarse-grained sampling step, guided by knowledge- or physics-based energy functions, the loop conformation space is explored to produce a large ensemble of reasonable, coarse-grained models satisfying geometric constraints. These coarse-grained models usually use reduced representations for loop structures, such as ϕ-ψ angles, backbone atoms, Cα atoms only [59], or side chain centers of mass [68]. Afterward, the coarse-grained models are filtered to eliminate the unreasonable ones in the ensemble [60] and then the representative models are selected by a clustering algorithm to reduce redundancy. These representative models are used to build fine-grained models in the refining phase, usually guided by a more accurate energy function associated with more structural information such as side chains and hydrogen atoms. Finally, in the ranking phase, the final models are assessed and the top-ranked ones are determined as the predicted results [62]. Among all these modeling steps, the coarse-grained sampling phase is of particular importance – if the sampling process cannot reach conformations close enough to the native, it is unlikely to obtain a high-resolution near-native model eventually. Moreover, the success of sampling relies on the underlying energy (scoring) functions, which are required to provide not only accurate, but also sensitive guidance to the sampling process to explore the protein loop conformation space.

There has been a lot of work done in modeling proteins loops since late 1960s. Limited by length, it is not our intention to provide a thorough review of loop modeling approaches in this mini review. Instead, we focus on the recent computational sampling approaches developed for protein loop structure prediction using template-free methods. We put our emphasis on the important factors that impact loop conformation sampling efficiency, including energy functions for modeling loops, loop buildup algorithms to satisfy geometric constraints, and coarse-grained sampling algorithms. The results of recent works in loop structure prediction are also summarized.

Energy Functions for Loop Modeling

According to Anfinsen's thermodynamics hypothesis [26], the native protein structure having the native structure conformation has the minimum Gibbs free energy of all accessible conformations. Similar to the general protein folding problem, many efforts of loop modeling focus on minimizing the protein potential energy described by physics-based energy functions. Zhang et al. [27] designed a simplified soft-sphere potential to fast construct loops. Cui et al. [28] developed a grid-based force field for their Monte Carlo (MC) sampling approach. More recent works take advantage of the existing force fields and solvent models popularly used in molecular simulation. Rapp and Friesner [29] and de Bakker et al. [30] used the AMBER [31] force field with a Generalized Born solvent model.

Spassov et al. [32] adopted the CHARMM [33] force field in their LOOPER algorithm. The Protein Local Optimization Program (PLOP) [34] developed by Jacobson et al. is based on the OPLS-AA [35] force field with Surface Generalized Born (SGB) solvent model. Rapp et al. [36] also used OPLS-AA/SGB to reproduce loop geometries in experimental solution structures. Zhu et al. [37] included a hydrophobic term in SGB solvation model and achieved accuracy improvement in long loops ranging from 11 to 13 residues. Felts et al. [38] incorporated Analytical Generalized Born plus Non-Polar (AGBNP) [39] implicit solvent model into PLOP. Sellers et al. [40] used MM/GBSA (Molecular Mechanics/Generalized Born Surface Area) energy [41] for loop refinement in comparative modeling. Danielson and Lill [80] also used MM/GBSA to study flexible loops interacting with ligands. Fogolari and Tosatto [42] took advantage of the concept of "colony energy" by considering the loop entropy, an important component in flexible loops, as part of the total free energy.

Since the main goal in loop structure prediction is to model loop conformation with high accuracy instead describing the underlying physics [47], an alternative approach to assess the correctness of a loop conformation is knowledge-based energy functions. The rationale of knowledge-based energy function is to obtain "pseudo energy" based on statistical preferences of conformations for different geometries as obtained from the database of known protein structures. Compared to the physics-based energy functions, the knowledge-based energy functions have several attractive advantages. First of all, the knowledge-based energy functions implicitly capture interactions that are difficult to model in physics-based energy functions. Secondly, the knowledge-based scoring functions usually do not require all atom information of the loops, which is ideal to rapidly generate coarse-grained models. Thirdly, the knowledge-based potentials tend to be "softer" to tolerate structural imperfection – allowing better handling of uncertainties and deficiencies of the computer generated models.

Sippl's potentials of mean force approach [43] is one of the most notable methods to obtain knowledge-based energy functions. According to the inverse-Boltzmann theorem, the knowledge-based energy potential $U(f)$ for a feature f is calculated as

$$U(f) = -kTln\frac{P_{obs}(f)}{P_{ref}(f)},$$

where k is the Boltzmann constant, T is the temperature, $P_{obs}(f)$ is the observed probability in the database of known structures, and $P_{ref}(f)$ is the probability of the reference state. Possible features to which a pseudo-energy term can be assigned include pairwise atom distances, torsion angles, amino acid contacts, side chain orientation, solvent exposure, or hydrogen bond geometry. For example, DFIRE [44] and DOPE [45] energy functions are built on the statistics of distance of pair-wise atoms. The dipolar DFIRE (dFIRE) [46] adds orientation-dependent terms to DFIRE by treating each polar atom as a dipole. Rata et al. [17] developed a statistical potential for loops based on adjacent ϕ-ψ pair distribution in the context of all possible combinations of local residue types. Liang et al. (OSCAR-loop) [64] optimized the knowledge-based potential for backbone torsion angles as Fourier series. Galaktionov et al. [65] designed penalty functions based on residue-residue contact map representations to model loops over 20 residues. Burke and Deane [69] calculated a sequence-based scoring function to estimate the compatibility of a sequence with a certain loop class.

In practice, physics- and knowledge-based energy terms are also often combined together to enhance the accuracy of the energy functions for loop prediction. Fiser et al. [47] used an energy function where stereochemical features are obtained from CHARMM-22 [33] force field while the non-bonded interactions, solvation, torsion angle preferences are derived from statistics. This energy function and the corresponding loop modeling method are adopted in the Modeller program. Rohl et al. [48] and later Mandell et al. [49] used the Rosetta scoring function [50], a hybrid scoring function which has demonstrated its effectiveness in CASP experiments. Xiang et al. [51] developed a combined energy function with force-field energy and RMSD (Root Mean Square Deviation) dependent terms, which is used in their LOOPY program.

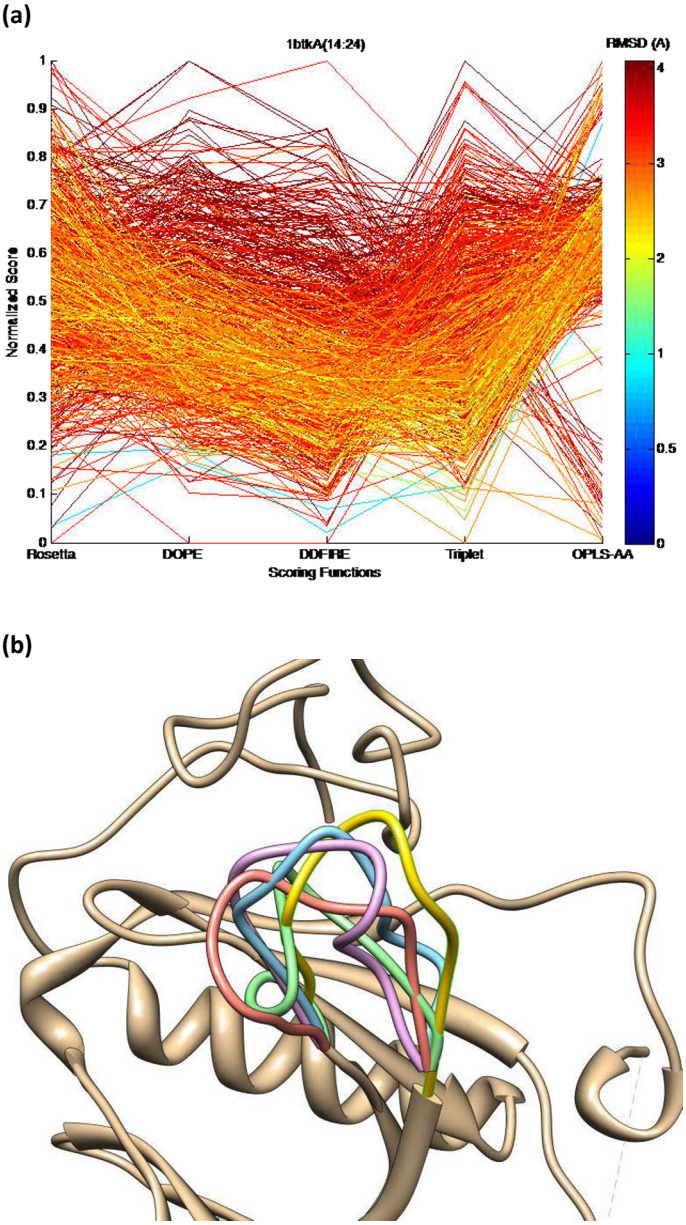

Figure 3. (a) Multiple energy functions coordinate plot of loop 1btkA(14:24) decoys in Jacobson loop decoy set using 5 energy functions, including Rosetta, DOPE, dDFIRE, backbone torsion potential using triplets, and OPLS-AA. All scores are linearly normalized in [0, 1]. RMSD is calculated for all backbone atoms in the loop. None of these energy functions can identify a near native decoy (< 1.0A) with the lowest energy value. **(b)** Native loop (gold) and loop decoys with lowest scores in Rosetta (blue, 2.73A), DOPE and dDFIRE (green, 2.85A), Triplet (red, 2.34A), and OPLS-AA (purple 2.27A).

Although quite a few energy functions derived by different manners are available for loop structure modeling, currently there does not exist a superiorly accurate energy function that can always differentiate the near native structures from the other incorrect ones in all protein loops. Figure 3(a) depicts a coordinate plot of multiple energy functions on decoys of 1btkA(14:24) contained in Jacobson loop decoy set [63] using a variety of physics-based, knowledge-based, or hybrid energy functions, including OPLS-AA [35], Rosetta [50], DOPE [45], dDFIRE [46], and backbone torsion potential using triplets [17]. None of these energy functions can correctly identify a near native decoy (< 1.0A) with the lowest energy value, although some near-native decoys exhibit low energy values in various energy functions. Figure 3(b) shows the loop decoy structures with the lowest energy values in different energy functions.

Loop Closure

The computer-generated loop models during the sampling process must satisfy the loop closure condition, i.e., the endpoints (C- and N-terminals) of a loop model must seamlessly bridge the anchored endpoints (C- and N-anchors) of the given protein structure. Figure 4 depicts the loop closure problem.

Computational methods enforcing loop models to satisfy loop closure constraints include energy penalty [52], finding analytical solutions [49, 53, 54], random tweak [55], wriggling [56], Cyclic Coordinate Descent (CCD) [57], or bi-directional inverse kinematics [58]. The energy penalty approach adds an additional term to the energy function to penalize deviations between the loop endpoints and the target anchor points [52]. The original method for obtaining analytical solutions is described in the pioneer work by Go and Scheraga [78]. Wedemeyer and Scheraga [53] derived the analytical solutions by determining the real roots of a polynomial, which lead to the solutions for closure of 6 backbone torsion angles in tripeptides. Coutsias et al. [54] and Mandell et al. [49] generalized the applicability of the analytical solutions to 6 not necessarily consecutive torsion angles in peptides of any length while small perturbations in bond angles and peptide torsion angles are also allowed. The random tweak method [55] is carried out by applying small random changes to ϕ-ψ angles and then using an iterated linearized Lagrange multiplier algorithm to satisfy the loop closure constraints with minimal conformational perturbations. Wriggling [56] takes advantage of the linear dependency of every four angles of rotation to keep the combined motion of loop localized. The CCD algorithm [57] treats the loop closure problem as an inverse kinematics problem, which fixes one loop endpoint at the one anchor and then iteratively modifies the ϕ-ψ angles in sequential order to minimize the distance between the other loop endpoint and the target anchor. The Full CCD (FCCD) algorithm [59] extends the applicability of CCD to a reduced loop representation with Cα atoms only by using a singular value decomposition-based optimization of a general rotation matrix. The bi-directional inverse kinematics method [58] adopts the "meet in the middle" strategy by generating half-loops from both C- and N-anchors and then assembles the endpoints of the half loops, which is particularly suitable for modeling long loops. In [60], Soto et al. provided a comparison of effectiveness and computational performance among various loop closure algorithms.

The above methods ensure loop closure, however, without considering the other geometric constraints such as steric clashes. Several methods have been proposed to account for additional geometric constraints in loop modeling. Xiang et al. [51] imposed a non-bonded energy term on the iterated Lagrange multiplier in the random tweak method to avoid steric clashes while satisfying loop closure simultaneously. Liu et al. [61] designed a self-organizing algorithm by performing fast weighted superimpositions of rigid fragments and adjusting distances between random atom pairs to resolve steric clashes, where not only loop closure, but also steric, planar, chiral, and even constraints derived from experiments can be satisfied simultaneously.

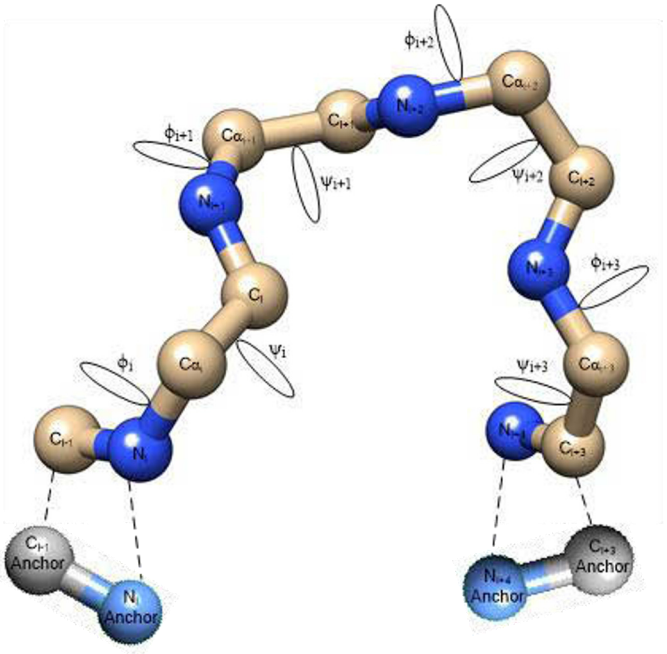

Figure 4. Addressing ϕ-ψ angles of a 4-residue loop to bridge the gap between the targeted anchored points

Loop Conformation Sampling

The loop conformation sampling is usually done by sampling backbone torsion angle conformations by deterministic or statistical sampling methods. In practice, it is not computationally feasible to sample all combinations of discretized torsion angles for a relatively long loop. Indeed, a large portion of these torsion angle combinations are infeasible due to steric clashes, unable to close, excluded volume for side chains, etc. In principle, both deterministic and statistical sampling techniques try to avoid these infeasible conformations as many as possible.

Deterministic sampling intends to find all possible loop conformations with reasonable but diversified structures. Galaktionov et al. [65] built loops (up to 12 residues) based on all possible combinations of local minima of empirical conformational energy for ϕ-ψ angles of each residue. Jacobson et al. [34] and Zhu et al. [37] developed rotamer libraries in PLOP for backbone torsion angles from high-resolution protein structure database. Then, loops are built up from the rotamer libraries while a variety of screening criteria, including effective resolution, clashes, impossible closure, deviation from protein body, and space for side chains, are used during sampling to eliminate as many infeasible structures as possible. Zhao et al. [79] extended the rotamer libraries to dipeptide segments to model long loops over 13 residues. Spassov et al. [32] performed a systematic search of ϕ-ψ angles belonging to one of the low energy basins in the iso-energy contour of local interactions.

Instead of attempting to generate all reasonable loop conformations, statistical sampling methods focus on the statistical favorability of conformations likely yielding low energy in the energy

Table 1. Energy functions, sampling methods, and loop closure mechanisms in recently published loop structure modeling works.

Loop Modeling Methods	Energy Functions	Coarse-grained Sampling	Fine-grained Sampling	Loop Closure
Fiser et al. [47] (2000) (Modeller)	Statistical potential integrating simple restraints or pseudo-energy terms	Random Buildup	Conjugate gradients – MD with simulated annealing – Conjugate gradients	Guaranteed in random buildup
Deane and Blundell [88] (2000) (PETRA) Deane and Blundell [76] (2001) (CODA)	Statistical all-atom, distance-dependent conditional probability function	Search Polypeptide Fragment Database	-	Filtering based on closure gap
Xiang et al. [51] (2002) (LOOPY)	Colony energy	Random Buildup	Fast torsion minimizer	Random tweak
DePristo et al. [66] (2003) De Bakker et al. [30] (2003) (RAPPER)	RAPDF-1 and RAPDF-2 (coarse) AMBER-GBSA (fine)	Sample dihedral angles from fine-grained torsion angle state sets	Limited-memory BFGS	Gap-closure restraint
Rohl et al. [48] (Rosetta) (2004) Mandell et al. [49] (2009) (Rosetta-KC)	Rosetta	MC, Simulated Annealing	MC energy minimization of all-atom Rosetta scoring function	Gap closure term in energy function [48] Kinematic closure [49]
Jacobson et al. [34] (2004) Zhu et al. [37] (2006) Zhao et al. [79] (2011) (PLOP)	OPLS-AA SGB (A hydrophobic term is added later in [37])	Rotamer Library Buildup	PLOP (Truncated Newton Local Optimization)	Meet in the middle
Spassov et al. [32] (2008) (Looper)	CHARMM with polar hydrogen force field parameters	Sampling backbone torsion angles in low energy basins of iso-energy contour	Newton-Raphson Minimization	Meet in the middle
Soto et al. [60] (2008) (Loop Builder)	DFIRE (coarse) OPLS/SBG-NP (fine)	Random sampling (same as LOOPY)	PLOP	Direct tweak
Cui et al. [28] (2008)	Grid-based force field	Local move MC, Simulated Annealing	Steepest Descent Energy Minimization	Filter local moves with reverse proximity criterion
Jamroz and Kolinski [74] (2010)	-	Hybrid Modeller, Rosetta, and CABS		-
Lee et al. [75] (2010)	DFIRE	Fragment Assembly	Side Chain Optimization	Analytical loop closure
Li et al. [72] (2011) (POS)	Rosetta, DFIRE, Triplet	MOMCMC	PLOP	CCD
Liang et al. [64] (2012) (OSCAR-loop)	Backbone potential, OSCAR force field, OPLS/SGB-NP	Random Buildup	Energy Minimization	CCD

landscape. In Modeller [47] and LOOPY [51], a lot of random loop conformations are generated and then optimized by energy minimization. In RAPPER [66], an ensemble of conformations with pair-wise RMSD greater than 0.2A is collected using a round-robin algorithm, in which a suitable ϕ-ψ combination satisfying geometric constraints is selected to gradually grow the loops. Lee et al. [75] produced loop conformations by sequentially adding randomly chosen 7-residue fragments obtained from known structure database. Ring and Cohen [70] sampled loop conformations with Genetic Algorithms (GA). More popularly, the Markov Chain Monte Carlo (MCMC) method [27, 28, 48, 49, 52, 61, 63, 64, 67, 68] is adopted to explore loop conformation space. The fundamental idea of MCMC is to perform local MC moves to propose new loop conformations satisfying loop closure and other geometric constraints without disturbing the rest of the protein structures and then decide the acceptance according to Metropolis acceptance-rejection criterion [71]. Various techniques have been used to enhance MC sampling efficiency, including simulated annealing [27, 28, 48, 49, 52, 64], hierarchical MC [63], replica exchange [67], and configuration-biased MC [68].

Generally, from algorithm point of view, GA is usually more effective than MC in terms of number of iteration steps to convergence, mainly due to better local minima escaping capability in GA when genetic operators such as crossover are employed [86]. However, in loop modeling, new conformations generated by crossover or mutation likely break the loop closure condition while potentially cause steric clashes. Additional quality control steps, potentially computationally costly, are necessary to correct these violations in geometric constraints [86]. In contrast, local MC moves in MCMC sampling guarantee satisfaction in geometric constraints and thus is more favorable in exploring loop conformation space.

After sampling, a set of coarse-grained loop models exhibiting good geometric properties are generated. Refining loop models, usually guided by a more accurate and sensitive energy potential

Table 2. Loop prediction accuracy in recently published works. The number of loop targets is specified in curly brackets.

Methods	Data Source	Average RMSD (A) {Number of Loop Targets}																
		Loop Length																
		2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18+
Fiser et al. [47] (2000) (Modeller)	Figure 9	0.5 {40}	0.5 {40}	1 {40}	1 {40}	2 {40}	2 {40}	2.5 {40}	3.5 {40}	3.5 {40}	5.5 {40}	6 {40}	6.5 {40}	6 {40}	-	-	-	-
Deane and Blundell [76] (2001) (CODA)	Table V	-	0.78 {153}	1.09 {184}	1.96 {162}	2.36 {97}	3.29 {78}	3.5 {60}	-	-	-	-	-	-	-	-	-	-
Xiang et al. [51] (2002) (LOOPY)	Table I	-	-	-	0.85 {161}	0.92 {107}	1.23 {74}	1.45 {61}	2.68 {58}	2.21 {34}	3.52 {37}	3.42 {21}	-	-	-	-	-	-
De Bakker et al. [30] (2003) (RAPPER)	Table III	0.35 {34}	0.37 {34}	0.47 {35}	0.9 {35}	0.95 {36}	1.37 {38}	2.28 {32}	2.41 {37}	3.48 {37}	4.94 {33}	4.99 {34}	-	-	-	-	-	-
Rohl et al. [48] (Rosetta) (2004)	Table II and Table VI	-	-	0.69 {40}	-	-	-	1.45 {40}	-	-	-	3.62 {40}		5.15 {avg. over 10 13- to 35-residue loops}				
Jacobson et al. [34] (2004) (PLOP)	Table IX	-	-	0.2 {35}	0.24 {117}	0.28 {100}	0.3 {82}	0.44 {66}	0.51 {57}	1.09 {40}	1.87 {18}	1.93 {10}	-	-	-	-	-	-
Zhu et al. [37] (2006) (PLOP)	Table II	-	-	-	-	-	-	-	-	-	1 {38}	1.15 {31}	1.25 {35}	-	-	-	-	-
Spassov et al. [32] (2008) (Looper)	Table I	0.26 {40}	0.31 {40}	0.42 {40}	0.49 {40}	0.81 {40}	1.07 {40}	1.33 {40}	1.63 {40}	2.66 {40}	3.35 {40}	4.08 {40}	-	-	-	-	-	-
Soto et al. [60] (2008) (Loop Builder)	Table V	-	-	-	-	-	-	1.31 {63}	1.88 {56}	1.93 {40}	2.5 {54}	2.65 {40}	3.74 {40}	-	-	-	-	-
Cui et al. [28] (2008)	Table I	-	-			0.75 {avg. over 14 4- to 9-residue loops}				-	-	-	-	-	-	-	-	-
Mandell et al. [49] (2009) (Rosetta-KC)	Figure 2	-	-	-	-	-	-	-	-	-	-	0.8 {63}	-	-	-	-	-	-
Jamroz and Kolinski [74] (2010)	Table II	-	-		1.07 {49}				2.23 {64}				-	-	-		7.87 {73}	
Lee et al. [75] (2010)	Table IV	-	-	0.54 {35}	0.92 {35}	1.36 {36}	1.17 {38}	1.87 {32}	2.08 {37}	3.09 {37}	3.43 {33}	3.84 {34}	-	-	-	-	-	-
Li et al. [72] (2011) (POS)	Tables II and III	-	-		0.33 {252}			0.58 {205}			0.86 {68}		0.63 {35}	-	-	-	-	-
Zhao et al. [79] (2011) (PLOP)	Table III	-	-	-	-	-	-	-	-	-	-	-	-	1.19 {36}	1.55 {30}	1.43 {14}	2.3 {9}	-
Liang et al. [64] (2012) (OSCAR-loop)	Table III	-	-	0.4 {2809}	0.52 {1863}	0.7 {1456}	0.83 {1053}	1.1 {862}	1.6 {634}	2.08 {528}	2.73 {392}	3.58 {325}	-	-	-	-	-	-

associated with more structural information such as side chain and hydrogen atoms, is needed to build fine-grained models. Similar to refining the complete protein structure, commonly used approaches to refine loop structures include local optimization [34], MC [81], and more often Molecular Dynamics (MD) simulations [82-85]. Furthermore, it is important to notice that coarse-grained and fine-grained sampling can be combined together to enhance exploration of loop conformation space, as an example shown in [67] where MC and MD simulations are integrated by a replica exchange algorithm.

Each loop modeling method has certain inevitable inaccuracy due to the limitation of sampling methods, uncertainty in energy functions, numerical errors, etc. A new strategy is to integrate different modeling methods to account for different sources of inaccuracy. Deane and Blundell [76] generated consensus predictions from two separate algorithms based on real fragments and computer generated fragments, respectively. Li et al. [72] developed a Pareto Optimal Sampling (POS) method based on the Multi-Objective Markov Chain Monte Carlo (MOMCMC) algorithm [73] to sample the function space of multiple knowledge- and physics-based energy functions to discover an ensemble of diversified structures yielding Pareto optimality. Jamroz and Kolinski [74] proposed a multi-method approach using MODELLER, Rosetta, and a coarse-grained de novo modelling tool, which leads to better loop models than those generated by each individual method.

Recent Loop Prediction Results

Table 1 summarizes the energy functions, sampling methods, and loop closure mechanisms and Table 2 lists the loop prediction accuracies in recently (since 2000) published works. Due to advances in computational loop modeling methods, highly accurate models with resolution comparable to experimental results have been achieved in quite a few methods shown in Table 2 for loops less than 8 residues. Several recent methods [37, 49, 72] can predict loop conformations within or close to 1A RMSD for loop targets up to 13 residues. Another important factor leading to loop modeling improvement is the stable growth of the number of known structures in PDB, which allows one to derive more sensitive knowledge-based energy functions, calibrate physics-based energy functions to achieve higher accuracy, and obtain richer loop fragments or rotamer libraries. Nevertheless, for the very long loops, for example, those over 18 residues, significant breakthrough has not been reported yet. According to Galaktionov et al. [65], modeling very long loops is a "different problem" due to their significantly higher flexibility compared to relatively short loops, which demands "different methodological approaches."

It is also important to notice that Table 2 does not serve the purpose of comparing prediction accuracy between different methods. First of all, the prediction accuracies in different methods are reported on different loop targets. Some loop targets are significantly "harder" than the others due to strong external influences from ions, ligands, disulfide bonds, and/or interactions with external chains or other units in the crystallographic unit cell. Several difficult loop targets (1poa(79:83), 1eok(A147:A159), 1hxh(A87:A99), and 1qqp(2_161:2_173)) are analyzed in [62]. Secondly, different criteria have been used to measure the accuracies of their prediction results in different methods. The RMSD calculations may be adopted very differently – either based on all heavy atoms, backbone atoms, or Cα atoms only. Moreover, the RMSD comparison may be directly carried out between the predicted model and the native structure, between the model and the relaxed native structure minimized by a force field, or between structures after global superimposition. Thirdly, loops are modeled under different assumptions in different methods. For example, Rosetta repacks all side chains of the protein [48, 49] while most of the other methods keep the native side chain conformations in the rest of the protein during the loop modeling process. Therefore, Table 2 does not form a fair base for comparing performance among different loop prediction methods, but is instead used to reflect the recent progress in loop modeling.

Summary

Loops play a critical role in performing important biological functions of proteins. However, due to their high flexibility and variability, modeling the 3D structures of loops is more difficult than other secondary structures. Loop structure modeling is regarded as a "mini protein folding problem" under geometric constraints such as loop closure and steric clashes. The computational loop modeling methods can be categorized into template-based and template-free methods. The template-based methods rely on database search, which is limited by the number of known structures in PDB, particularly when modeling relatively long loops. In comparison, the template-free methods can avoid this problem by diversely sampling loop conformation space to search for appropriate structures. Hence, sampling loop conformation space is the cornerstone of the template-free methods. Successful sampling methods rely on accurate and sensitive energy functions, fast buildup mechanism to generate reasonable loop models satisfying geometric constraints, and efficient sampling algorithms.

There has been remarkable advancement in template-free loop structure modeling in the past decade, mainly due to new computational methods as well as increasing number of known structures available in PDB. Quite a few loop modeling methods with various strategies have successfully predicted short loops (< 8 residues) with resolution comparable to experimental results. Several recent methods have even achieved near sub-angstrom accuracy in longer loops up to 13 residues. However, modeling very long loops over 18 residues is a challenge remaining unaccomplished. Recent study by Raval et al. [87] on protein structure refinement using very long (>100μs) MD simulations has shown that inaccuracy in current force fields limits MD-based protein structure refinement. Similarly, given loop modeling as a "mini protein folding problem," for difficult or long loop targets, while sampling is no longer a critical issue [87], development of more precise energy functions is now the key.

Acknowledgements

This work is partially supported by NSF grant 1066471, ODU 2011 SEECR grant, and ODU 2013 Multidisciplinary Seed grant.

Citation

Li Y (2013) Conformational Sampling in Template-Free Protein Loop Structure Modeling: An Overview. Computational and Structural Biotechnology Journal. 5 (6): e201302003. doi: http://dx.doi.org/10.5936/csbj.201302003

References

1. Steichen JM, Kuchinskas M, Keshwani MM, Yang J, Adams JA, Taylor SS (2012) Structural basis for the regulation of protein kinase A by activation loop phosphorylation. J Biol Chem 287: 14672-14680.

2. Ciarapica R, Rosati J, Cesareni G, Nasi S (2003) Molecular recognition in Helix-Loop-Helix and Helix-Loop-Helix-Leucine zipper domains. J Biol Chem 278: 12182-12190.
3. Bernstein LS, Ramineni S, Hague C, Cladman W, Chidiac P, Levey AI, Hepler JR (2004) RGS2 binds directly and selectively to the M1 muscarinic acetylcholine receptor third intracellular loop to modulate Gq/11alpha signaling. J Biol Chem 279: 21248-21256.
4. Kiss C, Fisher H, Pesavento E, Dai M, Valero R, Ovecka M, Nolan R, Phipps ML, Velappan N, Chasteen L, Martinez JS, Waldo GS, Pavlik P, Bradbury AR (2006) Antibody binding loop insertions as diversity elements. Nucl Acids Res 34: 132-146.
5. Stuart D, Acharya K, Walker N, Smith S, Lewis M, Phillips D (1986) Lactalbumin possesses a novel calcium binding loop. Nature 324: 84-87.
6. Saraste M, Sibbald PR, Wittinghofer A (1990) The P-loop: a common motif in ATP- and GTP-binding proteins. Trends Biochem Sci 15: 430-434.
7. Slesinger PA, Jan YN, Jan LY (1993) The S4-S5 loop contributes to the ion-selective pore of potassium channels. Neuron 11(4): 739-749.
8. Bruccoleri RE (2000) Ab initio loop modeling and its application to homology modeling. Methods in Molecular Biology 143: 247-264.
9. Dmitriev OY, Fillingame RH (2007) The rigid connecting loop stabilizes hairpin folding of the two helices of the ATP synthase subunit c. Protein Sci 16(10): 2118-2122.
10. Martin AC, Cheetham JC, Rees AR (1989) Modeling antibody hypervariable loops: a combined algorithm. Proc Natl Acad Sci USA 86(23): 9268-9272.
11. Espadaler J, Querol E, Aviles FX, Oliva B (2006) Identification of function-associated loop motifs and application to protein function prediction. Bioinformatics, 22(18): 2237-2243.
12. Tasneem A, Iyer LM, Jakobsson E, Aravind L (2005) Identification of the prokaryotic ligand-gated ion channels and their implications for the mechanisms and origins of animal Cys-loop ion channels. Genome Biol 6(1) R4.
13. Yarov-Yarovoy V, Baker D, Catterall WA (2006) Voltage sensor conformations in the open and closed states in ROSETTA structural models of K+ channels. Proc Natl Acad Sci USA 103: 7292–7297.
14. Flory PJ (1969) Statistical Mechanics of Chain Molecules. Wiley: New York.
15. Grdadolnik J, Grdadolnik SG, Avbelj F (2008) Determination of conformational preferences of dipeptides using vibrational spectroscopy. J Phys Chem B 112: 2712-2718.
16. Panchenko AR, Madej T (2005) Structural similarity of loops in protein families: toward the understanding of protein evolution. BMC Evol Biol 5: 10.
17. Rata I, Li Y, Jakobsson E (2010) Backbone Statistical Potential from Local Sequence-Structure Interactions in Protein Loops. J Phys Chem B 114(5): 1859-1869.
18. Kabsch W, Sander C (1983) On the use of sequence homologies to predict protein structure: Identical pentapeptides can have completely different conformations. Proc Natl Acad Sci USA 81: 1075-1081.
19. Cohen BI, Presnell SR, Cohen FE (1993) Origins of structural diversity within sequentially identical hexapeptides. Protein Sci 2: 2134–2145.
20. Fitzkee NC, Fleming PJ, Rose GD (2005) The Protein Coil Library: a structural database of nonhelix, nonstrand fragments derived from the PDB. Proteins 58(4): 852-854.
21. Wang G, Dunbrack RL (2003) PISCES: a protein sequence culling server. Bioinformatics 19(12): 1589-1591.
22. Greer J (1980) Model for haptoglobin heavy chain based upon structural homology. Proc Natl Acad Sci USA 77: 3393-3397.
23. Ring CS, Kneller DG, Langridge R, Cohen FE (1992) Taxonomy and conformational analysis of loops in proteins. J Mol Biol 224(3): 685-699.
24. Tramontano A, Lesk AM (1992) Common features of the conformations of antigen-binding loops in immunoglobulins and application to modeling loop conformations. Proteins, 13: 231-245.
25. Kwasigroch J, Chomilier J, Mornon J (1996) A global taxonomy of loops in globular proteins. J Mol Biol 259(4): 855-872.
26. Anfinsen CB (1973) Principles that govern the folding of protein chains. Science 181(4096): 223-230.
27. Zhang H, Lai L, Han Y, Tang Y (1997) A Fast and Efficient Program for Modeling Protein Loops. Biopolymers 41: 61-72.
28. Cui M, Mezei M, Osman R (2008) Prediction of Protein Loop Structures using a Local Move Monte Carlo Approach and a Grid-based Force Field. Protein Eng Des Sel 21(12): 729-735.
29. Rapp CS, Friesner RA (1999) Prediction of loop geometries using a generalized born model of solvation effects. Proteins 35: 173-183.
30. de Bakker PIW, Depristo MA, Burke DF, Blundell TL (2003) Ab initio construction of polypeptide fragments: accuracy of loop decoy discrimination by an all-atom statistical potential and the AMBER force field with the Generalized Born solvation model. Proteins 51: 21-40.
31. Cornell WD, Cieplak P, Bayly CI, Gould IR, Merz KM (1995) A second generation force-field for the simulation of proteins, nucleic-acids, and organic-molecules. J Am Chem Soc 117: 5179-5197.
32. Spassov VZ, Flook PK, Yan L (2008) LOOPER: a molecular mechanics-based algorithm for protein loop prediction. Protein Eng Des Sel 21(2): 91-100.
33. Brooks BR, Bruccoleri RE, Olafson BD, States DJ, Swaminathan S, Karplus M (1983) CHARMM: a program for macromolecular energy, minimization and dynamics calculations. J Comput Chem 4: 187-217.
34. Jacobson MP, Pincus DL, Rapp CS, Day TJF, Honig B, Shaw DE, Friesner RA (2004) A Hierarchical Approach to All-atom Protein Loop Prediction. Proteins 55: 351-367.
35. Damm W, Frontera A, Tirado-Rives J, Jorgensen WL (1997) OPLS All-Atom Force Field for Carbohydrates. J Comput Biol 18(16): 1955-1970.
36. Rapp CS, Strauss T, Nederveen A, Fuentes G (2007) Prediction of protein loop geometries in solution. Proteins 69: 69-74.
37. Zhu K, Pincus DL, Zhao S, Friesner RA (2006) Long Loop Prediction Using the Protein Local Optimization Program. Proteins 65: 438-452.
38. Felts AK, Gallicchio E, Chekmarev D, Paris KA, Friesner RA, Levy RM (2008) Prediction of Protein Loop Conformation Using the AGBNP Implicit Solvent Model and Torsion Angle Sampling. J Chem Theory Comput 4(5): 855-868.
39. Gallicchio E, Levy RM (2004) AGBNP: An Analytic Implicit Solvent Model Suitable for Molecular Dynamics Simulations and High-Resolution Modeling. J Comput Chem 25: 479-499.
40. Sellers BD, Zhu K, Zhao S, Friesner RA, Jacobson MP (2008) Toward better refinement of comparative models: predicting loops in inexact environments. Proteins 72(3): 959-971.
41. Kollman PA, Massova I, Reyes C, Kuhn B, Huo SH, Chong L, Lee M, Lee T, Duan Y, Wang W, Donini O, Cieplak P, Srinivasan J, Case DA, Cheatham TE (2000) Calculating structures and free energies of complex molecules: Combining molecular mechanics and continuum models. Accounts Chem Res 33(12): 889–897.

42. Fogolari F, Tosatto SCE (2005) Application of MM/PBSA colony free energy to loop decoy discrimination: toward correlation between energy and root mean square deviation. Protein Sci 14(4): 889-901.
43. Sippl MJ (1990) Calculation of conformational ensembles from potentials of mean force – an approach to the knowledge-based prediction of local structures in globular proteins. J Mol Biol 213: 859-883.
44. Zhou H, Zhou Y (2002) Distance-scaled, finite ideal-gas reference state improves structure-derived potentials of mean force for structure selection and stability prediction. Protein Sci 11: 2714–2726.
45. Shen M, Sali A (2006) Statistical potential for assessment and prediction of protein structures. Protein Sci 15(11): 2507-2524.
46. Yang Y, Zhou Y (2008) Specific interactions for *ab initio* folding of protein terminal regions with secondary structures. Proteins 72: 793-803.
47. Fiser A, Do RKG, Sali A (2000) Modeling of loops in protein structures. Protein Sci 9: 1753-1773.
48. Rohl RA, Strauss CE, Chivian D, Baker D (2004) Modeling structurally variable regions in homologous proteins with Rosetta. Proteins 55: 656-677.
49. Mandell DJ, Coutsias EA, Kortemme T (2009) Sub-Angstrom Accuracy in Protein Loop Reconstruction by Robotics-Inspired Conformational Sampling. Nat Methods 6: 551-552.
50. Simons KT, Bonneau R, Ruczinski I, Baker D (1999) Ab initio protein structure prediction of CASP III targets using Rosetta. Proteins 37(S3): 171-176.
51. Xiang ZX, Soto CS, Honig B (2002) Evaluating conformational free energies: the colony energy and its application to the problem of loop prediction. Proc Natl Acad Sci USA 99: 7432-7437.
52. Collura V, Higo J, Garnier J (1993) Modeling of protein loops by simulated annealing. Protein Sci 2: 1502-1510.
53. Wedemeyer WJ, Scheraga HA (1999) Exact Analytical Loop Closure in Proteins Using Polynomial Equations. J Comput Chem 20(8): 819-844.
54. Coutsias EA, Seok C, Jacobson MP, Dill KA (2004) A kinematic view of loop closure. J Comput Chem 25: 510-528.
55. Shenkin S, Yarmush DL, Fine RM, Wang H, Levinthal C (1987) Predicting antobody hypervariable loop conformation. I. Ensembles of random conformations for ringlike structures. Biopolymers 26: 2053-2085.
56. Cahill S, Cahill M, Cahill K (2003) On the kinematics of protein folding. J Comput Chem 24: 1364-1370.
57. Canutescu AA, Dunbrack RL (2003) Cyclic Coordinate Descent: A Robotics Algorithm for Protein Loop Closure. Protein Sci 12: 963-972.
58. Kolodny R, Guibas L, Levitt M, Koehl P (2005) Inverse Kinematics in Biology: The Protein Loop Closure Problem. Int J Robot Res 24: 151-163.
59. Boomsma W, Hamelryck T (2005) Full cyclic coordinate descent: solving the protein loop closure problem in Cα space. BMC Bioinformatics 6: 159.
60. Soto CS, Fasnacht M, Zhu J, Forrest L, Honig B (2008) Loop modelling: sampling, filtering, and scoring. Proteins 70: 834-843.
61. Liu P, Zhu F, Rassokhin DN, Agrafiotis DK (2009) A Self-Organizing Algorithm for Modeling Protein Loops. PLoS Comput Biol 5(8): e1000478.
62. Li Y, Rata I, Chiu S, Jakobsson E (2010) Improving Predicted Protein Loop Structure Ranking using a Pareto-Optimality Consensus Method. BMC Struct Biol 10: 22.
63. Nilmeier J, Hua L, Coutsias EA, Jacobson MP (2011) Assessing protein loop flexibility by hierarchical Monte Carlo sampling. J Chem Theory Comput 7(5): 1564-1574.
64. Liang S, Zhang C, Sarmiento J, Standley DM (2012) Protein loop modelling with optimized backbone potential functions. J Chem Theory Comput 8: 1820-1827.
65. Galaktionov S, Nikiforovich GV, Marshall GR (2001) Ab initio modelling of small, medium, and large loops in proteins. Peptide Science 60(2): 153-168.
66. DePristo MA, de Bakker PIW, Lovell SC, Blundell TL (2003) Ab initio construction of polypeptide fragments: efficient generation of accurate, representative ensembles. Proteins 51(1): 41-55.
67. Olson MA, Feig M, Brooks CL (2008) Prediction of protein loop conformations using multiscale modeling methods with physical energy scoring functions. J Comput Chem 29(5): 820-831.
68. Wu M, Deem MW (1999) Efficient Monte Carlo methods for cyclic peptides. Mol Phys 97: 559-580.
69. Burke DF, Deane CM (2001) Improved protein loop prediction from sequence alone. Protein Eng 14(7): 473-478.
70. Ring CS, Cohen FE (1994) Conformational sampling of loop structures using genetic algorithms. Israel Journal of Chemistry 34: 245–252.
71. Metropolis N, Rosenbluth AW, Rosenbluth MN, Teller AH, Teller E (1953) Equation of State Calculations by Fast Computing Machines. J Chem Phys 21: 1087-1092.
72. Li Y, Rata I, Jakobsson E (2011) Sampling Multiple Scoring Functions Can Improve Protein Loop Structure Prediction Accuracy. J Chem Inf Model 51(7): 1656-1666.
73. Li Y (2012) MOMCMC: An Efficient Monte Carlo Method for Multi-objective sampling over real parameter space. Comput Math Appl. 64: 3542-3556.
74. Jamroz M, Kolinski A (2010) Modeling of loops in proteins: a multi-method approach. BMC Struct Biol 10: 5.
75. Lee J, Lee D, Park H, Coutsias EA, Seok C (2010) Protein loop modeling by using fragment assembly and analytical loop closure. Proteins 78: 3428-3436.
76. Deane CM, Blundell TL (2001) CODA: a combined algorithm for predicting the structurally variable regions of protein models. Protein Sci 10: 599-612.
77. Regad L, Martin J, Nuel G, Camproux A (2010) Mining protein loops using a structural alphabet and statistical exceptionality. BMC Bioinformatics 11: 75.
78. Go N, Scheraga H (1970) Ring closure and local conformational deformations of chain molecules. Macromolecules 3: 178–187.
79. Zhao S, Zhu K, Li J, Friesner RA (2011) Progress in super long loop prediction. Proteins 79: 2920-2935.
80. Danielson ML, Lill MA (2012) Predicting flexible loop regions that interact with ligands: the challenge of accurate scoring. Proteins 80: 246-260.
81. Hu X, Wang H, Ke H, Kuhlman B (2007) High-resolution design of a protein loop. Proc Natl Acad Sci USA 104: 17668-17673.
82. van Schaik RC, Berendsen HJC, Torda AE, van Gunsteren WF (1993) A structure refinement method based on molecular dynamics in four spatial dimensions. J. Mol. Biol. 234: 751-762.
83. Hornak V, Simmerling C (2003) Generation of accurate protein loop conformations through low-barrier molecular dynamics. Proteins 51(4): 577-590.
84. Olson MA, Chaudhury S, Lee MS (2011) Comparison between self-guided Langevin dynamics and molecular dynamics simulations for structure refinement of protein loop conformations. J. Comp. Chem. 32(14): 3014-3022.

85. Dall'Agno KCM, de Souza ON (2012) An expert protein loop refinement protocol by molecular dynamics simulations with restraints. Expert Systems with Applications in press.
86. Unger R (2004) The genetic algorithm approach to protein structure prediction, Structure and Bonding 110: 153-175.
87. Raval A, Piana S, Eastwood MP, Dror RO, Shaw DE (2012) Refinement of protein structure homology models via long, all-atom molecular dynamics simualtions. Proteins 80: 2071-2079.
88. Deane CM, Blundell TL (2000) A novel exhaustive search algorithm for predicting the conformation of polypeptide segments in proteins. Proteins 40(1): 135-144.

Competing Interests:
The authors have declared that no competing interests exist.